Analysis of Criticisms of Validity Generalization Research and Selection Programs Based on Cumulative Validity Generalization Findings

Frank Schmidt, OPM and GWU
John Hunter, MSU
Kenneth Pearlman, OPM

Research evidence showing the generalizability of employment test validities has been well accepted scientifically and professionally. Nevertheless, some have expressed doubts and criticisms about this work. Some of these questions reflect genuine concerns and good faith efforts to obtain clarification on points not fully understood. Some criticisms appear not to have been advanced in good faith. Many questions of both kinds concern the legal defensibility of selection programs based on cumulative validity generalization research findings.

There has been only one court decision to date in which validity generalization findings (and the cumulative findings on differential validity and test fairness) have played a major role. This is the Pegues case. The judge accepted these findings and emphasized them heavily in his decision. Excerpts from that opinion are reproduced on the last page of this document. However, in this case, defendants had conducted their own studies (although they were conducted in other parts of the country and on other applicant populations), and so the defense did not rest entirely on validity generalization. There has not yet been a case in which the defense was based completely on validity generalization studies conducted by others. However, since employers are now basing selection programs on validity generalization findings by us and others, there may be such cases in the future. Thus it is important to examine questions of legal defensibility in addition to substantive and scientific questions.

This document presents a compilation of all the questions about, and objections to, validity generalization that we have been able to gather and outlines the answers to these objections as we see them.

Criticisms of, and Questions Raised About, Validity Generalization Research

1. There is nothing in the UGLs that allows for validity generalization; therefore such a procedure does not meet the UGLs requirements and should be struck down. (See p. 4 for answer.)

2. Because of reporting bias, nonsignificant and small validities will not be available--or will be disproportionately unavailable--and this unrepresentativeness of the data will cause validity generalization results to be biased, i.e., to overestimate validity. (See p. 9 for answer.)

3. Many of the studies used are very old and outdated; therefore the results are dubious. (See p. 12 for answer.)

4. There is no evidence that adequate job analyses were conducted in these original studies. (See p. 12 for answer.)

5. There is no evidence that the appropriate precautions against bias in criterion ratings were taken in the original studies. (See p. 12 for answer.)

6. The test fairness studies required by the UGLs were not conducted in the original studies. (See p. 13 for answer.)

7. Many of the original studies do not contain minorities and/or women, and thus are based on unrepresentative samples; this is particularly true of the older studies. (See p. 13 for answer)

8. The studies from which data was taken have not been shown to meet UGLs requirements, and, in all likelihood, do not meet these requirements. (See p. 13 for answer.)

9. Correcting for <u>both</u> criterion unreliability and range restriction is not psychometrically valid. See the 1974 APA <u>Standards</u> on this point. (See p. 13 for answer.)

10. Validity generalization is basically the process of averaging. Because of all this averaging, jobs for which the test has different validity or zero validity will be "swamped" by the rest of the data, leading to the false conclusion that the test is valid for those jobs. (See p. 14 for answer.)

11. Until the findings are replicated by others, validity generalization research findings and conclusions will remain somewhat suspect. As matters now stand, validity generalization findings are being reported only by one group of researchers - Schmidt, Hunter, Pearlman, and their associates. (In fact, other researchers do not even appear to employ their validity generalization data analysis methods.) In science, any research finding is justifiably suspect until it is replicated by others.

    This principle is especially important in this case since Schmidt, Hunter, and Pearlman are all employed by the U.S. Office of Personnel Management (the old Civil Service Commission), probably the largest user of employment tests in the world. It is interesting to note that their research findings are all consistent with the needs and policies of their employer. This does not mean their findings are false; it merely alerts us to the need for careful scrutiny of their research and the need for replication. (See p. 14 for answer.)

12. In their published papers, Schmidt, Hunter, and Pearlman - and also Callender and Osburn - do not present the observed validity coefficients from individual studies that form the basis of their validity generalization findings. This makes it impossible for others to replicate their analyses. Because of this, their findings published to date are suspect. (See p. 16 for answer.)

13. Even granting the scientific legitimacy of validity generalization findings to date, it would appear that in order to be really certain of conclusions about validity, the accuracy of the validity generalization conclusions should be empirically checked by any employer using them in his own setting. That is, validity generalization conclusions should be tested and verified by the employer's own empirical validity study. (See p. 16 for answer.)

14. Validity generalization conclusions are based on all information (i.e., previous validity studies) about a given test-job family combination that is available at the time of the validity generalization study. This may amount to, for example, 200 validity estimates. But the critical question is: what assurance is there that if 200 additional validity studies were conducted in the future, the results would be similar or comparable? It seems to me that we do not have this assurance, and therefore we cannot attribute very much credibility to validity generalization findings. (See p. 17 for answer.)

15. When empirical estimates of artifacts are not available, validity generalization results are based on assumed distributions of criterion reliability and range restriction effects. The researchers involved have maintained that these distributions are conservative; that is, they state that their assumed values of criterion reliability and range restriction (the restricted test standard deviation) are probably less variable than would be the case in reality. But this assertion cannot be proven. If they are wrong, and their distributions are more variable than reality, then they are attributing too much variance to artifacts, are overcorrecting, and are overestimating the generalizability of validities. We should therefore be skeptical of validity generalization findings. (See p. 18 for answer.)

16. It is best to be skeptical of validity generalization findings. This research and the methods it employs are still new and the whole thing is still controversial. There is no professional consensus within the field for acceptance of validity generalization. (See p. 18 for answer.)

17. Validity generalization must be false because it leads to the absurd conclusion that all cognitive aptitude or ability tests are equally valid for all jobs. (See p. 19 for answer.)

18. Acceptance of validity generalization would mean increased use of aptitude and ability tests, greatly reducing employment opportunities for members of minorities that have lower average test scores. (See p. 20 for answer.)

19. Acceptance of validity generalizability findings would produce a sort of Gresham's law: bad tests would drive out the good tests. Tests tailored to the specific requirements of the job would be replaced by general aptitude and ability tests, and we would be back to the bad old days of indiscriminate test use. (See p. 20 for answer.)

20. The fundamental concept of validity generalization is good but research to date has cumulated the wrong statistic. It is the regression of job performance on test scores that is important, not the correlation (validity) coefficient. Therefore, what should be cumulated and integrated across validation studies in validity generalization research is regression slopes (and intercepts), not correlation coefficients. Current validity generalization data should all be reanalyzed along these lines. (See p. 21 for answer.)

21. I don't object to the data reported on validity generalization; nor do I object to the methods of analysis. What I do object to is the strong statements and conclusions the Schmidt-Hunter-Pearlman group make. Instead of simply discussing their findings in the usual scientific manner, they are constantly making unduly strong conclusionary statements like "These results show that such-and-such is completely false" or "These findings show that . . . ." This tendency gives their work a quality of "flashiness", or "one-sidedness" that makes it hard to accept their findings. (See p.22 for answer.)

1. There is nothing in the UGLs that allows for validity generalization; therefore such a procedure does not meet the UGLs requirements and should be struck down.

ANSWER:

A. The UGLs do address validity generalization under the heading of transportability (Section 7; see also Section 8). Validity generalization research findings meet all the UGLs requirements for transportability, as shown below. When in litigation, this part of the defense should be presented at the beginning of the trial, in the testimony on direct by the defense expert witness, in order to establish on the record the scientifically correct interpretation of Section 7 of the UGLs. By anticipating the incorrect interpretation that plaintiffs will later attempt to force on this section, the effect of plaintiff's erroneous interpretation will likely be blunted.

Here is the text of Sections 7 and 8 of the UGLs:

Sec. 7. Use of other validity studies. A. Validity studies not conducted by the user. Users may, under certain circumstances, support the use of selection procedures by validity studies conducted by other users or conducted by test publishers or distributors and described in test manuals. While publishers of selection procedures have a professional obligation to provide evidence of validity which meets generally accepted professional standards (see section 5C above), users are cautioned that they are responsible for compliance with these guidelines. Accordingly, users seeking to obtain selection procedures from publishers and distributors should be careful to determine that, in the event the user becomes subject to the validity requirements of these guidelines, the necessary information to support validity has been determined and will be made available to the user.

B. Use of criterion-related validity evidence from other sources. Criterion-related validity studies conducted by one test user, or described in test manuals and the professional literature, will be considered acceptable for use by another user when the following requirements are met:

(1) Validity evidence. Evidence from the available studies meeting the standards of section 14B below clearly demonstrates that the selection procedure is valid;

(2) Job similarity. The incumbents in the user's job and the incumbents in the job or group of jobs on which the validity study was conducted perform substantially the same work behaviors, as shown by appropriate job analyses both on the job or group of jobs on which the validity study was performed and on the job for which the selection procedure is to be used; and

(3) Fairness evidence. The studies include a study of test fairness for each race, sex, and ethnic group which constitutes a significant factor in the borrowing user's relevant labor market for the job or jobs in question. If the studies under consideration satisfy (1) and (2) above but do not contain an investigation of test fairness, and it is not technically feasible for the borrowing user to conduct an internal study of test fairness, the borrowing user may utilize the study until studies conducted elsewhere meeting the requirements of these guidelines show test unfairness, or until suchtime as it becomes technically feasible to conduct an internal study of test fairness and the results of that study can be acted upon. Users obtaining selection procedures from publishers should consider, as one factor in the decision to purchase a particular selection procedure, the availability of evidence concerning test fairness.

C. Validity evidence from multiunit study. If validity evidence from a study covering more than one unit within an organization satisfies the requirements of section 14B below, evidence of validity specific to each unit will not be required unless there are variables which are likely to affect validity significantly.

D. Other significant variables. If there are variables in the other studies which are likely to affect validity significantly, the user may not rely upon such studies, but will be expected either to conduct an internal validity study or to comply with section 6 above.

Sec. 8. Cooperative studies.--A. Encouragement of cooperative studies. The agencies issuing these guidelines encourage employers, labor organizations, and employment agencies to cooperate in research, development, search for lawful alternatives, and validity studies in order to achieve procedures which are consistent with these guidelines.

B. Standards for use of cooperative studies. If validity evidence from a cooperative study satisfies the requirements of section 14 below, evidence of validity specific to each user will not be required unless there are variables in the user's situation which are likely to affect validity significantly.

Requirements of Section 7 of the UGLs .

7B(1): The study must meet the UGLs requirements for criterion-related studies (Section 14B). These requirements are:

14B1. The study should be technically feasible.

No problems here.

14B2. There must be a job analysis or review of job information.

This was done when the validity generalization study was carried out and when each contributing original study was done. This was done to at least the level of detail required (Schmidt, Hunter and Pearlman, 1981), as shown by the fact that the resulting

validities are very homogeneous (i.e., SD_ρ is low). The finding of a small value for SD_ρ shows that the job analyses were fully adequate for all the studies. If the job analyses in some or all of the studies had been inadequate, then either the criterion measures used would have been inappropriate or jobs would have been misclassified (included when they shouldn't have been or vice versa), or both these errors would have occurred. Either or both of these would have caused large variation in validities and thus a large value for SD_ρ (and therefore a failure of validity to generalize). Since SD_ρ is small, job analyses could not have been inadequate. Also, this homogeneity shows that the criterion measures in the original studies were comparable or similar in factor composition. This similarity indicates freedom from bias in what was measured, and is consistent with the research literature showing there is little if any bias in supervisory ratings.

14B3. Criterion measures should be adequate. Many original studies used ratings of overall job performance, which are specifically allowed in this section "where a study of the job shows it is an appropriate criterion." This "study" is based on professional judgment, and that judgment was made by the researchers. Other studies used multiple ratings summed into a composite. The high homogeneity of validities across all studies shows high similarity in what was rated. Either all the criterion measures are quite good or all are poor and are poor in exactly the same way. The latter is not plausible.

14B4. Representativeness of sample. This section defines representativeness in terms of race and sex, which have been shown not to moderate whether ability and aptitude tests are valid. The research literature on this point should be summarized in the employers validity generalization report. A representative sample is one that is representative of the applicant pool on variables that affect validity (see 1980 Div. 14 Principles). Therefore, all samples used were in fact representative. (See the material below indicating the UGLs were intended to be interpreted in light of current research findings.)

14B5. Statistical significance of validity. No problem here.

14B6. Operational use of selection procedure - not relevant here.

14B7. Overstatement of validity findings - not relevant.

14B8. Fairness - addressed separately under 7B(3).

7B(2): The jobs the study was based on and the jobs the results are generalized to must involve "The same major work behaviors, as shown by appropriate job analysis" on both sets of jobs. Major work behaviors are those that can affect or moderate test validity if different; otherwise, for purposes of a selection study, they are by definition not major; they are trivial and inconsequential. The homogeneity of validities in the validity generalization study demonstrates that the major work behaviors were the same in these jobs. Since the new job (to which we are generalizing the findings) is the same kind of job, (i.e., a sample job from the same job group), its major work behaviors are the same also.

Furthermore, psychological researchers always examine major work behaviors when conducting validity studies, to at least the level of detail required in validity generalization--the level of the whole job. For example, when studying clerical jobs, the researcher will always ascertain that a given job is a clerical job and not a pipefitter job. Correct identification and classification of the general nature or job family of the job is all that is required in validating cognitive ability or aptitude tests (Schmidt, Hunter, & Pearlman, 1980; Hunter, 1980b).

7B(3): Evidence of test fairness is required.

(1) The general research literature showing that aptitude and ability tests are fair to blacks and Hispanics satisfies this requirement. (The UGLs mention test publishers as a source of evidence on test fairness; the scientific research literature clearly has more credibility than test publishers.) The review of this literature should be included in the employer's validity generalization report (e.g., see Schmidt, Hunter, and Caplan, API Report, 1981).

(2) In addition to (1), this test fairness requirement goes into effect only if and when (a) it becomes technically feasible for the user to construct such a study or (b) such studies are conducted elsewhere. In the interim, the UGLs allow the test to be used.

Sections 7C and 7D and 8A and 8B provide additional support for validity generalization. These sections address "multiunit" and "cooperative" studies. Cooperative studies are specifically encouraged. Validity generalization is simply a method (the statistically optimal method) for analyzing data from "multiunit" studies, where the units are different organizations. Likewise, there is no essential difference between validity generalization studies and cooperative studies. Thus these sections make clear that the UGLs are intended to allow validity generalization under circumstances essentially identical to those of validity generalization studies.

B. Other Relevant Aspects of the UGLs.

1. The UGLs are intended to be consistent with professional standards. Section 5C reads:

C. Guidelines are consistent with professional standards. The provisions of these guidelines relating to validation of selection procedures are intended to be consistent with generally accepted professional standards for evaluating standardized tests and other selection procedures, such as those described in the Standards for Educational and

Psychological Tests prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education (American Psychological Association, Washington, D.C., 1974) (hereinafter "APA Standards") and standard textbooks and journals in the field of personnel selection.

Since validity generalization is accepted by professional standards (it is included in the 1980 Div. 14 Principles and will probably be in the revised APA Standards), and since it is included in journals and standard textbooks, it meets UGLs requirements.

2. UGLs were intended to be interpreted in light of changing circumstances. Section VII of the Federal Register introduction to the UGLs states:

> The bulk of the guidelines deals with questions such as those discussed in the above paragraphs. Not all such questions can be answered simply, nor can all problems be addressed in the single document. Once the guidelines are issued, they will have to be interpreted in light of changing factual, legal, and professional circumstances.

3. The UGLs expressly do not preclude the development and use of new validation methods. Section 14A states:

> Nothing in these guidelines is intended to preclude the development and use of other professionally acceptable techniques with respect to validation of selection procedures.

4. It is the intention of the present administration to revise the UGLs as soon as possible to be consistent with current research knowledge and professional practice. (See article in Washington Post, Aug. 14, 1981.) Such revision has been advocated to agency heads by the president of Div. 14 and by the Chair of the APA Committee on Tests and Assessments. Validity generalization and other new methods may well be explicitly recognized in the new UGLs.

2. Objection to Validity Generalization Study Results: The Hypothesis that Data Used in These Studies is Unrepresentative.

Some critics allege that the data used in validity generalization studies exclude a disproportionate number of low and/or nonsignificant observed validities. These allegations may be phrased approximately as follows:

1. "Validity studies that do not result in positive, significant validity coefficients typically go unreported."

2. "There is an obvious bias against reporting non-significant validity coefficients. These studies are often not even written up, and if they are they will soon be lost or forgotten."

It is important to establish immediately that this allegation is an hypothesis or speculation - not a fact - and therefore the evidence for and against it must be carefully examined. In a court case, it might be best for the defense to introduce this hypothesis on direct examination of their expert witness and have him or her present the evidence against this hypothesis. This should take most of the impact out of the allegation when plaintiffs introduce it later. Consider the evidence that such a bias does not exist, i.e., the evidence that careful searches for unpublished and published studies, such as those that we and others have conducted, will produce a representative group of validity coefficients:

1. We considered this question in Schmidt and Hunter (1977). We quote:

> The next question that must be addressed in connection with this model is the possibility that validity coefficients available for inclusion in any given prior may not be representative of the relevant population of coefficients. This question is easily defined if priors are composed only of coefficients from the published literature. Wouldn't it be reasonable to assume that researchers are more likely to submit for publication, and editors are more likely to accept, studies reporting statistically significant and/or large validities than those reporting small and/or nonsignificant coefficients? The present study and those we have underway are based on unpublished as well as published coefficients. But this question may not disappear simply as the result of including unpublished studies, because the proportion of coefficients in the prior from unpublished studies may be less than the proportion in the coefficient population from unpublished studies. This outcome would be probable if published studies were more often located by the researchers than unpublished studies.
>
> There are other considerations, however, that indicate that this effect may be of little or no consequence. First, a major, and perhaps the major, source of validity coefficients is the Validity Information Exchange, which was published by Personnel Psychology between 1954 and 1965. The editorial policy of the Validity Information Exchange--which was made known to the journal's readers--was to accept submissions without regard to statistical significance of findings. Indeed, the editors make a special point of urging that studies showing nonsignificant results be submitted (Ross, 1961; Taylor, 1953). Of the 1,506 coefficients reported over the lifetime of the Exchange, 856 or 57% were nonsignificant (Lent, Aurbach, & Levin, 1971a).

A second consideration is that most studies contain not one but a number of validity coefficients, making it highly probable that at least some will be statistically significant. For example, if statistical power in a study reporting 10 independent coefficients is only .50, the then probability that at least one coefficient will be significant is .999. The probability that at least two will be significant is .989. For three, the probability is .95, and for four, it is .83. On the other hand, the probability that all will be nonsignificant is $.5^{10}$, which is essentially zero. What this means is that almost all studies will tend to have some significant coefficients, decreasing the probability that the researcher will base his decision to submit, or the editor his decision to accept, on statistical significance.

But perhaps the most important consideration is the fact that if such a selective process were operating, the studies it would screen out would tend to be the methodologically poorer ones, that is, those containing information of poorer quality to begin with. Studies reporting all or almost all nonsignificant and/or near zero validities will tend disproportionately to be those characterized by low statistical power--resulting, in turn, from low criterion reliabilities, use of small samples, and high levels of range restriction (cf. Schmidt, Hunter, & Urry, 1976). (In this connection, it is interesting to note that Boehm, 1977, found a significant negative relationship between methodological quality of studies and probability of reporting findings of single-group validity. Single-group validity has repeatedly been shown to be a chance phenomenon; see Hunter & Schmidt, 1978, for a review of this research.) Thus studies reporting nonsignificant results may tend to be rejected not so much on that basis as on the basis of methodological weaknesses. Methodological weaknesses are, after all, not invisible to reviewers. Those studies--probably small in number--that are methodologically sound but nevertheless report only nonsignificant coefficients may have good probabilities of being accepted for publication.

If the above hypothesis is true, what would be the effect of somehow being able to retrieve the rejected low quality studies and include their coefficients in the prior? The effect would be a decrease in the accuracy of the prior unless compensating adjustments were made in the assumed means and distributions of criterion reliabilities and range restriction effects and in the assumed average sample size. Specifically, the new distribution of criterion reliabilities would have to have a lower mean and greater variance. The new distribution of range restriction effects would have to show greater mean range restriction and more variance in range restriction effects. And the assumed mean sample size would have to be smaller. The result would be that larger corrections would be made to the mean and variance of the prior to allow for the lower quality of the input data.

2. Supporting the hypothesis that low and nonsignificant observed validities, when they occur, are due to methodological deficiencies in their respective studies, rather than to low true validities, is the finding from the data in Lent et al. (1971), that the correlation between percent of validity coefficients significant and median sample size was .77. The larger the sample size, the more likely was a finding of significant validity.

3. Where it is possible to compare data sets like that of Pearlman et al. (1980) to data sets known to be complete, the two data sets are very similar. For example, the Pearlman et al. data set is very similar to the U.S. Department of Labor GATB data set used by Hunter (1981) in terms of means and variances (controlling for sample sizes) of observed validity coefficients. The same is true when the comparison is with large sample military data sets; military researchers routinely report all data. Also, our mean observed rs are virtually identical to Ghiselli's reported medians, based on decades of careful information gathering.

4. In our experience examining many hundreds of unpublished studies, we have found it is not true that data is suppressed or omitted. Such studies virtually always report results for all tests tried out (even poorly constructed experimental instruments with well below average reliabilities). In the typical scenario, the study is an exploratory one designed to determine the optimal test battery. A multi-test battery is tried out on a variety of jobs and/or criteria. Full tables of validities against all criteria for all jobs are reported. Test batteries are then recommended based on regression analyses. There is thus no evidence that reporting the full set of results (including usually many low and nonsignificant rs) is viewed negatively by the sponsoring organization. We know of no evidence that organizations are willing to send substantial sums on validity studies and then allow or encourage either in-house psychologists or outside consultants to partially or fully suppress the results.

5. Even a cursory examination of validity generalization data sets indicates nonselectivity. For example, in the Pearlman et al. (1980) data set for proficiency criteria, 349 of the 2,795 observed validities (12.5%) are zero or negative. 737 (26.4%) were .10 or less. Further, 56.1% of the 2,795 observed validities were nonsignificant at the .05 level. This figure is consistent with our estimate (Schmidt, Hunter, and Urry, 1976) that the average criterion related validity study has statistical power no greater than .50. If selectivity or bias in reporting were operating, the percent significant would almost certainly have been much higher than 43.9% (because a disproportionate number of the nonsignificant rs would have been omitted).

6. However, the most striking fact is the close comparability of the percent of observed validities that were nonsignificant in the published studies reviewed by Lent et al. (1971) - 57% - and the percent nonsignificant in the mostly unpublished data set (68% unpublished) of Pearlman et al. (1980) - 56.1%. The Pearlman et al. data almost perfectly match the published data, a clear empirical indication that unpublished data is not different from published data.

7. The file drawer analysis of Rosenthal (1978) can be used to determine the number of unlocated studies not finding validity that would be necessary to change conclusions about validity. When file drawer analyses have been done on the results of validity generalization studies, the required number of missing studies has typically been so large as to have little possibility of existing (e.g., 200 to 10,000 or more). For a published example, see Callender and Osburn (1981). We would be glad to send copies of two of our unpublished reports (for Sears and for the American Petroleum Institute) that obtain similar results.

These findings cast doubt on the reporting bias hypothesis. In view of the above, the hypothesis of data suppression is not tenable.

3. Many of the studies used are very old and outdated; therefore the results are dubious.

ANSWER:

There is no evidence that validities of aptitude and ability tests change with time. In fact, the evidence shows almost perfect stability. Pearlman (1980), for example, has shown that the observed validities of tests used in clerical selection have remained virtually identical from the 1920s to the 1980s--despite whatever changes in job content and work force composition may have transpired over this period of almost half a century. If plaintiffs want to advance the hypothesis that test validities change or have changed over time, it is up to them to present evidence supporting this hypothesis. Otherwise, in the face of the evidence to the contrary, their position is merely a speculation--and a counterfactual one at that.

4. There is no evidence that adequate job analyses were conducted in these original studies.

ANSWER:

Yes, there is. First researchers virtually always analyze jobs to at least the level of detail required in validity studies of aptitude and ability tests. (See attack No. 2, above.) Second, the finding of a small value for SD_ρ shows that the job analyses were fully adequate for all the studies. If the job analyses in some or all of the studies had been inadequate, then either the criterion measures used would have been inappropriate or jobs would have been misclassified (included when they shouldn't have been or vice versa), or both these errors would have occurred. Either or both of these would have caused large variation in validities and thus a large value for SD_ρ (and therefore a failure of validity to generalize). Since SD_ρ is small, job analyses could not have been inadequate.

5. There is no evidence that the appropriate precautions against bias in criterion ratings were taken in the original studies.

ANSWER:

Yes, there is. (See also the reply to question No. 1 above.) Again, the small value for SD_ρ means that there was great similarity in what was measured by the criteria in the different studies. Either all the criterion measures are quite free of bias in what they measure or all are quite biased in what they measure--and based in exactly the same way. The latter is very implausible.

6. The test fairness studies required by the UGLs were not conducted in the original studies.

ANSWER:

a. See discussion of UGLs Section 7B (3) in reply to question No. 1, above.

b. Further, such studies were not technically feasible in all but all small percentage of the original studies, because of either the absence of minorities or insufficient sample sizes for both groups. The UGLs require fairness studies only if and when they are technically feasible.

7. Many of the original studies do not contain minorities and/or women, and thus are based on unrepresentative samples; this is particularly true of the older studies.

ANSWER:

See discussion of UGLs section 14B (4) in reply to question No. 1, above.

8. The studies from which data was taken have not been shown to meet UGLs requirements, and, in all likelihood, do not meet these requirements.

ANSWER:

This criticism is very general; this objection will have to go beyond this generality and specify the specific ways that the original studies do not meet UGLs requirements. These more specific criticisms are rebutted in the replies to questions 1, 3, 4, 5, 6, and 7, above.

9. Correcting for both criterion unreliability and range restriction is not psychometrically valid. See the 1974 APA Standards on this point.

ANSWER:

Use of both corrections is endorsed by professional standards (see 1980 Div. 14 Principles, pp. 10-11). The 1974 APA standards were in error on this point.(They are now under revision.) In the process of publishing Schmidt, Hunter, and Urry (1976), we convinced the editor of Journal of Applied Psychology and the three very critical reviewers of our paper of this fact. They had initially objected to use of both corrections, basing their objections on the 1974 APA Standards. If plaintiffs nevertheless object to the use of both corrections, it is up to them to provide reasons and evidence to support their position. Otherwise it is merely unfounded personal opinion (and again, this fact should be clearly stated in defense testimony). We have provided detailed psychometric reasons why both corrections are necessary in order to avoid biased validity estimates (Schmidt, Hunter, and Urry, 1976). In the five years since 1976, no one we have encountered who objected to the use of both corrections has been able to articulate any statistical or psychometric reasons for their objection. Thus it is critical for the defense to see to it that the burden of supporting these objections be placed squarely on the plaintiffs. Doing so will neutralize this attack.

10. Validity generalization is basically the process of averaging. Because of all this averaging, jobs for which the test has different validity or zero validity will be "swamped" by the rest of the data, leading to the false conclusion that the test is valid for those jobs.

ANSWER:

This argument is false for two reasons.

(1) If validities had been zero (or even widely different) for some of the jobs, the standard deviation of validities (SD_ρ) would have been much larger than in fact it is. The SD is very sensitive to outliers (much more so than the mean). And, in fact, we have found large SD_ρ values for very heterogeneous predictors such as performance tests (cf. Pearlman et al., 1980). Thus the small SD_ρ value indicates there were no such outliers.

(2) The process of setting confidence bounds on the true validity is the same process used in all of inferential statistics. For example, suppose the employer conducted his own validity study and obtained a statistically significant validity coefficient ($p = .05$). This finding merely means that the lower end of the 90% confidence interval around the observed validity does not include zero. It does not mean the true value cannot be zero; it means only that the confidence interval probably includes the true value--exactly the same principle used in validity generalization studies. Thus this argument is not an attack on validity generalization, it is a (false) attack on statistical methods in general--on grounds that since they do not provide 100% certainty, they should be discarded in favor of some alternative procedure. In validation, the alternative is the single employer-conducted validity study, which, as we have shown, provides much lower levels of confidence and certainty. (See also the answers to questions 13 and 14, below.)

11. Until the findings are replicated by others, validity generalization research findings and conclusions will remain somewhat suspect. As matters now stand, validity generalization findings are being reported only by one group of researchers - Schmidt, Hunter, Pearlman, and their associates. (In fact, other researchers do not even appear to employ their validity generalization data analysis methods.) In science, any research finding is justifiably suspect until it is replicated by others.

This principle is especially important in this case since Schmidt, Hunter, and Pearlman are all employed by the U.S. Office of Personnel Management (the old Civil Service Commission), probably the largest user of employment tests in the world. It is interesting to note that their research findings are all consistent with the needs and policies of their employer. This does not mean their findings are false; it merely alerts us to the need for careful scrutiny of their research and the need for replication.

ANSWER:

It is not the case that the only group conducting validity generalization research is the Schmidt-Hunter-Pearlman group. Many other individuals and groups are conducting research in this area and are obtaining comparable results. Callender and Osburn (1980) have derived their own equation for

estimating the standard deviation of true validities and have conducted computer simulation studies that show that their equation and our equations are quite accurate. They have also applied their validity generalization equations to empirical data, and, like us, have reached the conclusion that employment test validities are generalizable (Callender and Osburn, 1981). Working at the Life Insurance Management Research Association (LIMRA), Brown (in press) has applied validity generalization methods to LIMRA's Aptitude Index and has found generalizability. Linn, Hamish and Dunbar (1980), working at the University of Illinois, have applied validity generalization methods to correlations between LSAT scores and first year law school GPA from a large sample of law schools; again, the conclusion was that validity was generalizable. Working with employment tests, Timmreck (1981) reached the same conclusion at the University of Houston. Marvin Dunnette has conducted a large-scale test of validity generalization methods and predictions as part of a nationwide consortium validation study. Validity was found to be generalizable, and all predictions from earlier validity generalization studies (e.g., on the role of sampling error in producing variance in observed validities) were upheld. This study will be available in written form in the future, and will be prepared for publication. Using slightly different methods, Sharf (1981) has demonstrated the generalizability of validity for an index of managerial potential. Finally, research on validity generalization is underway in the CIA, at Armco, Inc., at Proctor and Gamble, and quite likely at other organizations which have not informed us of their activities. Thus validity generalization findings are not confined to one group of researchers but are quite widespread.

We have noted an apparent tendency on the part of opponents of validity generalization to attempt to impugn our credibility as scientists because two of us (Schmidt and Pearlman) are employed by the U.S. Office of Personnel Management. (John Hunter is a professor of psychology and mathematics at Michigan State University. He is not employed by OPM.) The insinuation that a scientist's work is or will be biased by reason of employment is as distasteful to us as it is to other industrial/organizational psychologists, whether employed in industry or government. Research questions and scientific issues cannot be decided by innuendo and ad hominem attacks; they must be resolved by careful, reasoned examination of the empirical research evidence. We believe the evidence we have presented for our conclusions is very strong. In addition, as noted above, our findings have been replicated by other researchers not associated with OPM.

The insinuation that our research findings have been biased in the direction of OPM needs or policies is rich with irony. Most of the research referred to was conducted and published during the Carter years (1977-81). The policies of the Carter administration were strongly anti-testing, and these policies were forcefully represented by Carter's Director of OPM, Alan Campbell. (One of Campbell's major achievements in the selection area was the phasing out of the nationwide PACE written exam by means of a sweetheart consent decree agreement with a group of plaintiffs.) Because our research findings and recommendations were very much opposed to administration policy, we were constantly concerned that management might intervene and cancel the research program. We received admonishments about the need to support official policy which sounded threatening. However, there was never any direct interference with our research. In light of OPM's opposition to, and distaste for, our research findings, it is indeed ironic to hear suggestions that pressures from OPM have biased our findings in the direction of favoring strong validity generalization conclusions.

12. In their published papers, Schmidt, Hunter, and Pearlman - and also Callender and Osburn - do not present the observed validity coefficients from individual studies that form the basis of their validity generalization findings. This makes it impossible for others to replicate their analyses. Because of this, their findings published to date are suspect.

ANSWER:

It is true that we have typically not published the observed validity coefficients that are the basis of our validity generalization studies. However, this has been the case only because of journal page limitations. Researchers interested in replicating any of our analysis may obtain the observed coefficients by writing to us. (Note: The thrust of this charge is greatly blunted anyway by the fact noted in (11), above, that other researchers using their own data have obtained similar validity generalization results.)

13. Even granting the scientific legitimacy of validity generalization findings to date, it would appear that in order to be really certain of conclusions about validity, the accuracy of the validity generalization conclusions should be empirically checked by any employer using them in his own setting. That is, validity generalization conclusions should be tested and verified by the employer's own empirical validity study.

ANSWER:

This suggestion is analogous to checking the accuracy of the powerful telescopes used in astronomy by looking at the night sky with the naked eye. The measure or method used as the evaluation standard should be less, not more, susceptible to error than the procedure being evaluated. Consider the amount of error in the typical employer-conducted validity study. Lent et al. (1971a; 1971b) found that the median sample size for published validity studies was only 68; unpublished studies may average lower. A study in this sample size range will usually have less than a 50-50 chance of detecting validity given that it is present (Schmidt, Hunter, & Urry, 1976). Even worse, in a sample size this small, the presence of even one outlier can change the observed validity from positive to negative. Suppose, for example, that the best worker is sick on the day the tests are administered but comes to work despite his illness, takes the tests, and scores uniformly low. He or she then has a very high criterion score and very low predictor scores. Small samples are so unstable that only one such case is required to produce the false appearance of a negative validity. Even without outliers, the 95% confidence interval around the observed validity (for N = 68) is 48 correlation points wide; that is, its width is .48. Small sample studies contain very little information and therefore provide very uncertain validity estimates.

Now consider the estimate of validity that is to have its accuracy checked against this "criterion." Estimates of validity from validity generalization studies are typically based on 100-200 validity coefficients (the minimum in our studies has been eight; the maximum, over 800). These coefficients, in turn, are based on a total sample size typically in the 8,000-20,000 range. Thus the data base for the validity generalization estimate of validity is massive; the amount of information in this data base is great, and the resulting estimate is very accurate. Even relatively large-sample single studies (e.g., N = 500 - 1500), in the rare cases in which they are possible, do not provide as sound a basis for validity estimation as the typical validity generalization study.

14. Validity generalization conclusions are based on all information (i.e., previous validity studies) about a given test-job family combination that is available at the time of the validity generalization study. This may amount to, for example, 200 validity estimates. But the critical question is: what assurance is there that if 200 additional validity studies were conducted in the future, the results would be similar or comparable? It seems to me that we do not have this assurance, and therefore we cannot attribute very much credibility to validity generalization findings.

ANSWER:

The basis of this objection is doubt about the fundamental principle of induction in science. It is not, therefore, an objection specifically to validity generalization. Science progresses through induction of general laws from specific instances of their apparent manifestation. For example, once the gravitational constant had been measured numerous times at various points on the surface of the earth (all yielding to some value to within measurement error), induction led to the general principle that this parameter had the same value at all points of the planet's surface. Enough data have accumulated that scientists have confidence in this general principle--even though there are many parts of the planet where the gravitational constant has never been calibrated. The same principle applies in validity generalization; a point is reached at which certainty is great enough that gathering additional data would not be a profitable way to expend time and effort.

Also, it is important to note that in the case of most validity generalization findings, even if all the studies were conducted again and even if all the new studies obtained validity estimates of exactly zero, such a bizarre outcome would not be sufficient to change the original conclusion that the test is valid. In such a case, the best estimate of validity would be based on the two sets of studies combined. The file drawer analysis of Rosenthal (1978) would typically show that the two sets of studies combined would reach some substantial level of statistical significance. This conclusion follows from the fact that file drawer analyses almost invariably show that the number of unlocated studies with zero validity that must exist to bring the combined significance level of the present set of studies (based on observed r's) down to the .05 level is much greater than the number of studies in the present set.

Finally, it should be noted that while this objection is not valid against validity generalization findings, it is valid against the alternative to validity generalization, namely the single in situ employer-conducted validity study. That is, this objection implicitly suggests that a validity study conducted in the setting at hand would be superior to reliance on validity generalization findings. But, as described in the answer to (13) above, such single studies have such low statistical power and such wide confidence intervals for validity that a repetition of such a study--even in the same setting--is highly likely to produce a different result and a different conclusion. Unlike cumulations of many studies, single validity studies are highly unlikely to replicate. (For an empirical demonstration of this fact, see Bender and Loveless, Personnel Psychology, 1958, 11, 491-508.)

15. When empirical estimates of artifacts are not available, validity generalization results are based on assumed distributions of criterion reliability and range restriction effects. The researchers involved have maintained that these distributions are conservative; that is, they state that their assumed values of criterion reliability and range restriction (the restricted test standard deviation) are probably less variable than would be the case in reality. But this assertion cannot be proven. If they are wrong, and their distributions are more variable than reality, then they are attributing too much variance to artifacts, are overcorrecting, and are overestimating the generalizability of validities. We should therefore be skeptical of validity generalization findings.

ANSWER:

This question was raised by those reviewing our papers for publication several years ago. In response, we reanalyzed our data dropping all assumed distributions of artifacts (criterion and test reliabilities and range restriction effects). In these analyses, we corrected the variance of observed validities only for the effects of sampling error. (The exact amounts of sampling error expected are known from the standard formulas for the sampling error of a correlation coefficient.) In every case, conclusions about the generalizability of validity were unchanged. Even the specific numerical results were almost the same. We have referred to this simplified analysis as "bare bones" validity generalization and have used our findings with this method to show that the generalizability of validities could (and should) have been "discovered" as early as the 1920s or 1930s. In general, about 75% - 85% of the variance in observed validity coefficients for job performance that artifacts account for is accounted for by simple sampling error. Other artifacts are relatively unimportant in comparison to sampling error. Psychologists have been underestimating the effects of sampling error for at least six decades. For further details and validity generalization analyses based only on corrections for sampling error, see Pearlman et al. (1980) and Schmidt, Gast-Rosenberg, and Hunter (1980).

16. It is best to be skeptical of validity generalization findings. This research and the methods it employs are still new and the whole thing is still controversial. There is no professional consensus within the field for acceptance of validity generalization.

ANSWER:

The first thing to be noted about this objection is that it is not substantive. That is, it is not addressed to the substantive merits of validity generalization research and findings. There is no allegation that such research is methodologically or otherwise flawed.

Second, validity generalization is no longer "new." It has been around since 1976. Quite a number of psychologists have done research on validity generalization, and their results are in agreement in finding that validity is generalizable. See the response to (11), above, for details.

Third, validity generalization does enjoy professional acceptance. It has been recognized in the 1980 revision of the Division 14 Principles for Validation and Use of Employee Selection Procedures. But perhaps more importantly, this research has been repeatedly accepted by and published in the two major scientific journals in the field - the Journal of Applied Psychology JAP, and Personnel Psychology (PP). This means the research has been thoroughly and carefully

scrutinized by numerous reviewers especially chosen by the journal editors for their psychometric and statistical expertise. These reviews have been extremely thorough and detailed, especially those done for JAP. (There were at least three and typically four reviewers for each JAP validity generalization paper.) In addition, the editors themselves carefully examined each paper. Acceptance from, and publication in, these journals constitutes a high level of scientific and professional acceptance. The fact that the editors of the October 1981 Special Issue on Testing of the American Psychologist invited us to write the article on research findings in employment testing is another indication of professional acceptance. Thesame applies to the invitation received to write the chapter on employment testing for the Annual Review of Psychology. The many invitations to make presentations on validity generalization research at conferences and universities are another such indication. We have also received many recommendations from colleagues that validity generalization findings be incorporated into the revised Federal Uniform Guidelines on Employee Selection Procedures and the revised APA-AERA-NCME Standard for Psychological Tests. Finally, in light of the number of personnel researchers (mostly I/O psychologists) who have adopted validity generalization for use in their organizations, the statement also appears to be false. Psychologists in the following organizations have introduced validity generalization as the basis for selection programs:

1. 13 large petroleum companies
2. U.S. Government (as the basis of all entry-level clerical selection and for certain other selection programs)
3. Armco Inc. (basis of clerical selection)
4. U.S. Employment Service (basis of all GATB testing nationwide)
5. CIA
6. Sears

In short, we have every reason to be gratified at the professional and scientific acceptance that has been accorded our research findings.

17. Validity generalization must be false because it leads to the absurd conclusion that all cognitive aptitude or ability tests are equally valid for all jobs.

ANSWER:

Validity generalization research leads to no such conclusion. In fact our research has demonstrated that there are reliable differences between jobs in the validity of most aptitude and ability tests (Schmidt and Hunter, 1978; Schmidt, Hunter, and Pearlman, 1981).

However, the fact that any given test is more valid for some jobs than for others does not mean that the test is invalid for any job. In fact our research provides strong evidence that reliable measures of the standard aptitudes (e.g., verbal, quantitative, and spatial abilities) are valid predictors of both performance in training and performance on the job for all jobs in the occupational spectrum. To many, this sounds like a radical conclusion, and relative to past beliefs, it is. However, the evidence exists to support this conclusion. Hunter (1980; "Validity generalization for 12,000 jobs. . .") examined a representative sample of over 500 jobs covering the entire occupational spectrum from highest to lowest level jobs and showed that the abilities measured by the Labor Department's General Aptitude Test Battery (GATB) were valid for predicting both job performance and training success in all of these

jobs. Because this sample of jobs is representative of the occupational spectrum, these findings can be generalized to all jobs in the economy for which tests are used in selection. Hunter grouped jobs into five job families varying in job complexity, and showed that these families moderated validity in the limited sense that given abilities have higher validity for some job families than for others. However, all abilities have some (substantial) degree of validity for all job families. Since this massive study is critical to the broadest conclusion yet reached about the generalizability of validity, we suggest that all who are seriously interested in validity generalization obtain a copy from John Hunter and read it carefully. Copies can also be obtained from the U.S. Office of Personnel Management and from the U.S. Employment Service.

18. Acceptance of validity generalization would mean increased use of aptitude and ability tests, greatly reducing employment opportunities for members of minorities that have lower average test scores.

ANSWER:

Acceptance of validity generalization would mean wider use of aptitude and ability tests, but this would not necessarily mean increased adverse impact or reduced employment opportunities for minorities. In fact, it could mean reduced adverse impact. A recent quantitative meta-analysis of alternatives to written tests by Hunter has shown that most such alternatives have substantial adverse impact. Validity generalization might lead to replacement of these methods by ability tests. In our publications, we have described in detail how ability tests can be used in selection in such a way as to not merely reduce but eliminate adverse impact, while at the same time preserving 85%-90% of the productivity gains from valid selection. The critical point is that the question of adverse impact is separate from the question of validity generalization and/or test use, and can be manipulated and controlled independently. In fact, our experience indicates that many companies already do this.

19. Acceptance of validity generalizability findings would produce a sort of Gresham's law: bad tests would drive out the good tests. Tests tailored to the specific requirements of the job would be replaced by general aptitude and ability tests, and we would be back to the bad old days of indiscriminate test use.

ANSWER:

This objection is based on the assumption that tests that measure general aptitudes (e.g., spatial ability) are "bad" tests. In view of research showing that such tests have substantial levels of validity and can produce large increases in workforce output (Schmidt, Hunter, McKenzie & Muldrow, 1979), it is hard to see why they are "bad." Another assumption appears to be that tests that mirror the specific activities of the job are "good" tests. Perhaps here the implication is that they are more valid than aptitude tests. If so, we know of no empirical evidence to support that proposition. Even if it were true, evidence would still be needed to show that the validity of "job specific" tests did not disappear as job content changed over time. And, of course, the cost of constructing "job specific" tests for each job should not be overlooked. This cost would be prohibitive for most organizations.

20. The fundamental concept of validity generalization is good but research to date has cumulated the wrong statistic. It is the regression of job performance on test scores that is important, not the correlation (validity) coefficient. Therefore, what should be cumulated and integrated across validation studies in validity generalization research is regression slopes (and intercepts), not correlation coefficients. Current validity generalization data should all be reanalyzed along these lines.

ANSWER:

There are serious statistical problems standing in the way of cumulating regression slopes (or intercepts) across studies. Different studies use different scales for measuring job performance, and also different tests for measuring any given ability (e.g., numerical computation). Thus the regression slopes, $\beta = r_{xy}\ (SD_y / SD_x)$, are not comparable from study to study because SD_y and SD_x are not comparable. Regression slopes therefore cannot be cumulated across studies. If one attempts to solve this problem by standardizing criterion and test scores within each study, then the resulting beta weight is equal to r_{xy}, and one is back to the correlation or validity coefficient. The advantage of correlation coefficients is that they are in the same units across studies and can therefore be cumulated across studies. The advantage of regression slopes is that their use eliminates the need to correct for range restriction and criterion unreliability, since these two artifacts do not bias estimates of raw score regression slopes. Since such corrections magnify the effects of sampling error, they should be used only when they are needed. In the case of the correlation coefficient, they are needed if one is to avoid biased estimates of validity.

It is also important to note that the raw score slopes (and intercepts) are not the slope values needed to determine the practical value of an employment test to the using organization. The Brogden-Cronbach utility equation is the regression of dollar value of employee output onto test score (Schmidt, Hunter, McKenzie & Muldrow, 1979). If test score is standardized to N(0,1), this slope is $r_{xy_T}\ SD_y/1 = r_{xy_T}\ SD_y$, where SD_y = the standard deviation in dollars of employee output in a randomly selected (unrestricted) group, and r_{xy_T} is the true or operational validity. In employment testing, this is the slope value that is needed, and the one that should be cumulated across studies. We have cumulated r_{xy_T} across studies to produce more accurate estimates of the true validity, and have estimated SD_y separately for each job group. However, if the data were available (they aren't, to our knowledge), the product of these two $r_{xy_T}\ SD_y$, could be cumulated.

21. I don't object to the data reported on validity generalization; nor do I object to the methods of analysis. What I do object to is the strong statements and conclusions the Schmidt-Hunter-Pearlman group make. Instead of simply discussing their findings in the usual scientific manner, they are constantly making unduly strong conclusionary statements like "These results show that such-and-such is completely false" or "These findings show that . . . ." This tendency gives their work a quality of "flashiness" or "one-sidedness" that makes it hard to accept their findings.

ANSWER:

This objection questions our style of stating research findings and conclusions. It does not question the actual findings and conclusions. When there is a massive amount of empirical evidence supporting a conclusion, and there is little or no empirical evidence to the contrary (e.g., as with test fairness or validity generalization), the "strong" statements of conclusion which this objection correctly states that we make are not only appropriate but in fact are scientifically mandated. They are not "flashy" or "one-sided." Under these circumstances, weak wishy-washy conclusionary statements are in fact scientifically irresponsible--because they distort the known facts. Further, we believe that one (of many) reasons for the current meataxe slashing of social science research support by the Federal government is the (correct) public perception that most behavioral and social scientists are extremely reluctant, no matter how massive the empirical evidence, to make any straightforward conclusionary statements. Instead all the public gets for its tax dollars are weak, wishy-washy, squeamish statements encrusted with qualifications and hedged all around with reservations and self-doubt. There are many research areas and research conclusions for which all this is not only completely unnecessary, but misleading, but few in which these barnacles have been stripped off. The development of meta-analysis will aid in this process. (See Bibliography.)

Style of expression, in science as elsewhere, is to a great extent a matter of personal preference and personal style. Some people are timid and reservation-prone by nature, and show these traits even when they are clearly contraindicated by the known facts. This again points up the importance of focusing attention away from such matters and onto the empirical evidence. We believe that even people who, at the emotional level, do not like our style of communicating research findings will accept our conclusions if they can be induced to carefully examine and evaluate the empirical research evidence we present in support of them. In the long run, evidence is what counts, not style.

(Incidentally, if one looks back at the history of science, one finds that the great contributors virtually never spoke or wrote in a hedged-about over-qualified way. They were the first to clearly see the nature of phenomena, and they expressed their conclusions in clear, strong statements. Examples: Darwin, Mendlev (the periodic table of elements), Newton, Einstein. There is little support in the history of science for a squishy, squeamish, style of communication.)

Excerpts on Validity Generalization and Related Issues from Pegues, et al. v. Mississippi State Employment Service et al., 22FEP3929 (Northern District of Mississippi, March 7, 1980)

P. 48. Empirical research has demonstrated that validity is not perceptibly changed by differences in location, differences in specific job duties or applicant populations. Valid tests do not become invalid when these circumstances change. Plaintiffs' allegation that validity is specific to a particular set of tasks and to a specific applicant population, or in other words, that a valid test in one set of circumstances is not valid in circumstances not perfectly identical is not true. (Testimony of Dr. John Hunter.) 1/

(Footnote) 1/ The research done by Dr. John Hunter and his colleagues shows that the validity of tests is broadly generalizable across jobs, geographical locations and applicant populations. In both the 1970 EEOC Guidelines and the Standards of the American Psychological Association such research has been called for.

P. 50. No differences between the job duties in the research sample and the jobs in Bolivar County were specified. According to research, even gross changes in job duties did not destroy validity. It follows that small and/or hypothesized differences have little or no effect on validity. Plaintiffs have not shown that the USES tests were invalid because the tasks of the jobs in the research setting may have been different from those in Bolivar County. (Testimony of Dr. John Hunter.)

P. 52 Section 1607.4(c)(1) has been met. Differences in location, jobs and applicant characteristics do not alter the proper interpretation of validity evidence. Further, no credible evidence exists that differences between units, jobs and applicant populations alter test validity. (Testimony of Dr. Hunter.)

P. 53. Section 1607.5(b)(1) has been met. No evidence was presented that any of the samples are atypical or unrepresentative. Differences in applicant characteristics do not moderate or change validity. (Testimony of Dr. John Hunter.)

Bibliography

I. General Meta-Analysis Methods

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G. V. Meta-Analysis in Social Research. Beverly Hills, Calif.: Sage Publications, 1981.

Hunter, J. E., & Schmidt, F. L. Cumulating findings across studies: Correction for sampling error, a proposed moratorium on the significance test, and critique of current multivariate reporting practices. Unpublished paper, 1981.

Hunter, J. E., Schmidt, F. L., & Jackson, G. Quantitative Methods for Integrating Findings Across Studies: Advanced Meta-Analysis, Including Corrections for Distortions Due to Sampling Error, Measurement Error, and Range Restriction. Beverly Hills, Calif.: Sage Publications, in press.

II. Overviews of the State of the Art in Personnel Selection

Schmidt, F. L., & Hunter, J. E. New research findings in personnel selection: Myths meet realities in the 1980's. Public Personnel Administration: Policies and Procedures for Personnel. Englewood Cliffs, N.J.: Prentice-Hall, 1981.

Schmidt, F. L., & Hunter, J. E. Employment testing: Old theories and new research findings. American Psychologist, in press. (Special Issue on Testing, October 1981.)

III. Studies of Validity Generalization

Brown, S. H. Validity generalization in the life insurance industry. J. Appl. Psychol., in press.

Callender, J. C., & Osburn, H. G. Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results of petroleum industry validation research. Journal of Applied Psychology, 1981, 66, 274-281.

Callender, J. C., & Osburn, H. G. Development and test of a new model for validity generalization. Journal of Applied Psychology, 1980, 65, 543-558.

Hunter, J. E. Validity Generalization and Construct Validity. In Construct Validity in Psychological Measurement: Proceedings of a Colloquium on Theory and Application in Education and Measurement. Princeton, N. J., Educational Testing Service, 1980(a)

Hunter, J. E. Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). U.S. Employment Service, U.S. Department of Labor. Washington, D.C., 1980(b)

Hunter, J. E. An analysis of validity, differential validity, test fairness, and utility for the Philadelphia Police Officer Selection Examination Prepared by Educational Testing Service. Unpublished paper, Department of Psychology, Michigan State University, 1980.

Lilienthal, R. A., & Pearlman, K. The validity of Federal selection tests for aid/technicians in the health, science, and engineering fields (Personnel Research Report). U.S. Office of Personnel Management, Personnel Research and Development Center, in press.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. Unpublished manuscript, University of Illinois at Champaign-Urbana, Department of Psychology, 1980.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. Journal of Applied Psychology, 1980, 65, 373-406.

Pearlman, K., & Schmidt, F. L. Effects of alternate job grouping methods on selection procedure validity. In E. L. Levine (chair), Job Analysis/Job formulas: Current perspectives on research and application. Symposium presented at the meeting of the American Psychological Association, Los Angeles, August 24, 1981.

Schmidt, F. L., Hunter, J. E., & Urry, V. M. Statistical power in criterion-related validity studies. Journal of Applied Psychology, 1976, 61, 473-485.

Schmidt, F. L., & Hunter, J.E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schmidt, F. L., & Hunter, J. E. Moderator research and the law of small numbers. Personnel Psychology, 1978, 31, 215-231.

Schmidt, F. L., Hunter, J. E., Pearlman, K. & Shane, G. S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 1979, 32, 257-281. (a)

Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. Validity generalization results for computer programmers. Journal of Applied Psychology, 1980, 65, 643-661.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Caplan, J. R. Validity generalization results for three occupations in the Sears Roebuck Company. Chicago: Sears Roebuck Company, 1981. (99 pp.)

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. Validity generalization results for two jobs in the petroleum industry. Journal of Applied Psychology, 1981, 66, 261-273.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. Selection procedure validity generalization (transportability) results for three job groups in the petroleum industry. Washington, D.C.: The American Petroleum Institute, 1981. (130 pp.)

Schmidt, F. L., Hunter, J. E., & Pearlman, K. Task differences and validity of aptitude tests in selection: A red herring. Journal of Applied Psychology, 1981, 66, 166-185.

Sharf, J. C. Recent developments in the field of industrial and personnel psychology. Paper presented at the conference, "Recent Directions in Testing and Fair Employment Practices." Washington, D.C. Sponsored by the Personnel Testing Council of Metropolitan Washington and BNA Systems, April 23, 1981.

Timmreck, C. W. Moderating effect of tasks on the validity of selection tests. University of Houston, Department of Psychology, 1981.

IV. Single Group Validity, Differential Validity, and Test Fairness

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannon, R. Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 1978, 31, 233-241.

Boehm, V. R. Differential prediction: A methodological artifact? Journal of Applied Psychology, 1977, 62, 146-154.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. An investigation of sources of bias in the prediction of job performance. A six year study. Final Project Report No. PR-73-37. Princeton, N.J.: Educational Testing Service, 1973.

Gael, S., & Grant, D. L. Employment test validation for minority and non-minority telephone company service representatives. Journal of Applied Psychology, 1972, 56, 135-139.

Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority clerks with work sample criteria. J. Applied Psychology, 1975, 60, 420-426(a).

Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology, 1975, 60, 411-419(b).

Grant, D. L., & Bray, D. W. Validation of employment tests for telephone company installation and repair occupations. Journal of Applied Psychology, 1970, 54, 7-14.

Hunter, J. E., & Schmidt, F. L. A critical analysis of the statistical and ethical implications of five definitions of test fairness. Psychological Bulletin, 1976, 83, 6, 1053-1071.

Hunter, J.E., Schmidt, F. L., & Rauschenberger, J. M. Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 1977, 62, 245-260.

Hunter, J. E., & Schmidt, F. L. Differential and single group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Psychology, 1978, 63, 1-11.

Hunter, J. E., Schmidt, F. L., & Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. Methodological and statistical issues in the study of bias in mental testing. Chapter to appear in, Reynolds, C. R., Perspectives on Bias in Mental Testing. New York: Plenum Press, in press.

Katzell, R. A., & Dyer, F. J. Differential validity revived. Journal of Applied Psychology, 1977, 62, 137-145.

Ledvinka, J. The statistical definition of fairness in the Federal selection guidelines and its implications for minority employment Personnel Psychology, 1979, 32, 551-562.

Linn, R. L. Single-group validity, differential validity, and differential predictions. Journal of Applied Psychology, 1978, 63, 507-514.

O'Connor, E. J., Wexley, K. N., & Alexander, R. A. Single-group validity: Fact or fallacy? Journal of Applied Psychology, 1975, 60, 352-355.

Ruch, W. W. A re-analysis of published differential validity studies. Paper presented at the symposium, "Differential Validation Under EEOC and OFCC Testing and Selection Regulations." American Psychological Association. Convention, Honolulu, Hawaii, September 6, 1972.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. Personnel Psychology, 1980, 33, 705-724.

Tenopyr, M. L. Race and socio-economic status as moderators in predicting machine-shop training success. Paper presented at the 75th Annual Convention of the American Psychological Association, Washington, D.C., 1967.

V. Studies of Selection Utility Using New Methods

Hunter, J. E., & Schmidt, F. L. Fitting people to jobs: Implications of personnel selection for national productivity. Chapter to appear in E. A. Fleishman (Ed.), Human Performance and Productivity, 1981, in press.

Hunter, J. E. Fairness of the General Aptitude Test Battery (GATB): Ability differences and their impact on minority hiring rates. Report prepared for U.S. Employment Service, 1981.

Hunter, J. E. The economic benefits of personnel selection using ability tests: A state of the art review including a detailed analysis of the dollar benefit of U.S. Employment Service placements and a critique of the low cutoff method of test use. Report prepared for U.S. Employment Service, U.S. Department of Labor, Washington, D.C., January 15, 1981.

Hunter, J. E., & Schmidt, F. L. Noncompensatory aspects of personnel selection. Unpublished paper, 1980.

Mack, M. J., Schmidt, F. L., & Hunter, J. E. Estimating the productivity costs in dollars of minimum selection test cutoff scores. Personnel Research and Development Center, 1980.

Rauschenberger, J. The utility of valid selection procedures: Dollars and sense. Paper presented at the conference, "Recent Directions in Testing and Fair Employment Practices." Washington, D.C. Sponsored by the Personnel Testing Council of Metropolitan Washington and BNA Systems, April 23, 1981.

Schmidt, F. L., Hunter, J. E., McKenzie, R. & Muldrow, T. The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 1979, 64, 609-626. (b)

Schmidt, F. L., & Hunter, J. E. Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. Under review by Journal of Applied Psychology.

Schmidt, F. L., Hunter, J. E., Outerbridge, A., & Trattner, M. The economic impact of job selection methods on the size, productivity, and payroll costs of the Federal workforce: An empirical demonstration. 1981.

Schmidt, F. L. An empirical analysis of the economic impact of the Luevano Consent Decree on the size, productivity and payroll costs of the GS 5-7 Federal workforce. Personnel Research and Development Center, U.S. Office of Personnel Management, 1981.

Schmidt, F. L. The impact of the Maintenance-Helper selection test batteries of the Philadelphia Electric Company on workforce productivity and payroll costs. Unpublished paper, 1981.

Schmidt, F. L. The impact of the clerical employees selection test battery of the Philadelphia Electric Company on clerical productivity and payroll costs. Unpublished paper, 1981.

VI. Other:

Lent, R. H., Aurbach, H. A., & Levin, L. S. Predictors, criteria and significant results. Personnel Psychology, 1971, 24, 519-533.

Lent, R. H., et al. Research design and validity assessment. Personnel Psychology, 1971, 24, 247-274.

Pearlman, K. Seeing the whole picture: Application of cumulated validity data to issues in clerical selection. In V. J. Bentz (Chair), Methodological implications of large-scale validity studies of clerical occupations. Symposium presented at the meeting of the American Psychological Association, Montreal, September 2, 1980.